THE ROYAL STATISTICAL SOCIETY

2009 EXAMINATIONS – SOLUTIONS

ORDINARY CERTIFICATE

PAPER II

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

Ordinary Certificate, Paper II, 2009. Question 1

(i) Since all the data values are non-negative, the lowest possible value of each is 0 and therefore the lowest possible value of the mean is 0.

This occurs when all the data values are 0.

(ii) The lowest possible value of the variance is 0.

This occurs when all the data values are equal.

(iii) The lowest possible value of Spearman's rank correlation coefficient is –1.

This occurs when the two sets of rankings are completely reversed, i.e. $\{1, 2, \ldots, n-1, n\}$ and $\{n, n-1, \ldots, 2, 1\}$.

(iv) The lowest possible value of the product-moment correlation coefficient is –1.

This occurs when the values of the two variables lie on a straight line of negative slope.

Ordinary Certificate, Paper II, 2009. Question 2

The percentages in each category are rounded to the nearest whole number, so rounding errors have occurred.

Advantages of bar charts:
- They are easy to draw without calculation and with ruler only
- It is easy to compare directly on a scale the value of each category

Advantages of pie charts:
- They show the size of each category in relation to the whole
- They are visually appealing

The angle for "daytime cost" is $(61/99) \times 360 = 222°$ to the nearest whole number.

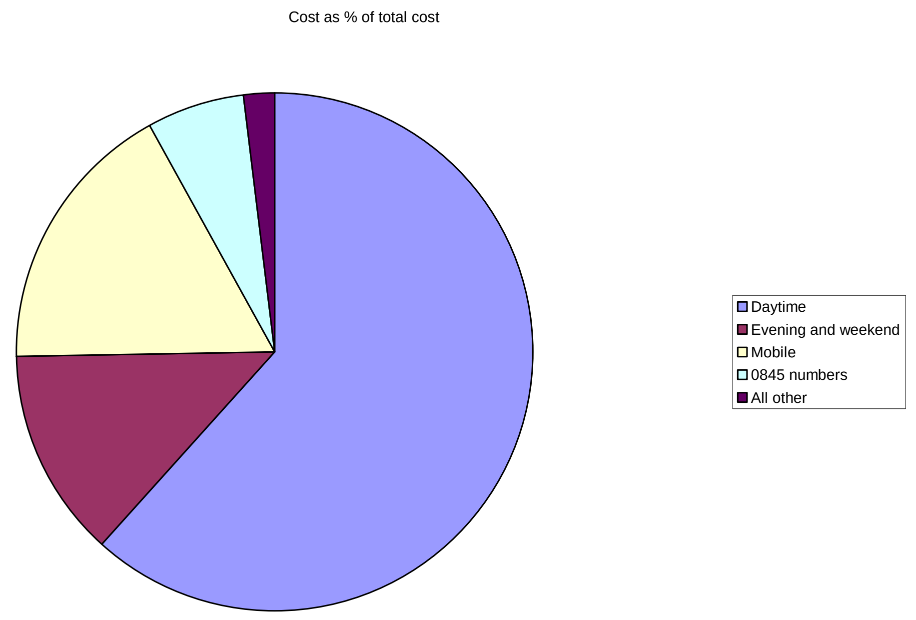
Thus, arguing similarly for the others, the angles are as follows.

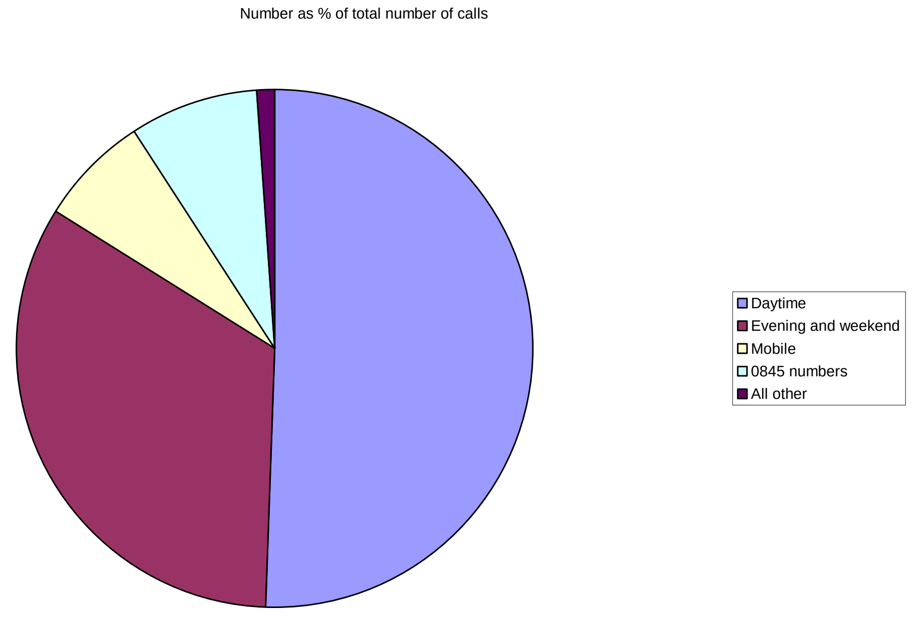
Type of call	Cost as % of total cost	Angle for pie chart
Daytime	61	222
Evening and weekend	13	47
Mobile	17	62
0845 numbers	6	22
All others	2	7
Total	99	360

Type of call	Number as % of total number of calls	Angle for pie chart
Daytime	50	182
Evening and weekend	33	120
Mobile	7	25
0845 numbers	8	29
All others	1	4
Total	99	360

The pie charts are shown on the next page.

Solution continued on next page





[Note. The use of colour is not necessary provided the sectors are clearly identifiable and labelled.]

Solution continued on next page

Comments might include the following.

The most common calls are daytime ones.

Daytime calls account for half of the calls but approaching two-thirds of the cost.

Evening and weekend calls account for one-third of the calls but a much smaller proportion of the cost.

The cost of mobile calls as a proportion of the total cost is more than double the proportion of mobile calls as a proportion of all calls.

Daytime and mobile calls are relatively more expensive than evening and weekend calls.

Ordinary Certificate, Paper II, 2009. Question 3

The data arranged in order of magnitude are shown in the table.

Numbers of calls in ascending order	
Centre A	Centre B
503	418
508	436
509	455
518	518
521	519
529	523
546	527
546	546
554	558
564	571
574	572
582	601
583	615
591	623
592	667

Range for centre A is 592 – 503 = 89 calls.
Range for centre B is 667 – 418 = 249 calls.

There are 15 days, so the median is the number of calls on the (15+1)/2 = 8th day when the daily calls are arranged in order.

Thus the median for centre A is 546 calls and the median for centre B is 546 calls.

There are 15 observations. The lower quartile is the number of calls on the (15+1)/4 = 4th day.

Similarly, the upper quartile is the number of calls on the 3×(15+1)/4 = 12th day.

[Note. Other conventions exist for defining the lower and upper quartiles. These were acceptable in the examination.]

Thus

lower quartile for centre A is 518 calls
upper quartile for centre A is 582 calls

lower quartile for centre B is 518 calls
upper quartile for centre B is 601 calls

Solution continued on next page

For both centres, on half the days the number of daily calls is 546 or fewer. The range of calls per day is higher (249) for centre B than for centre A (89). Although for both centres on a quarter of the days the number of calls is 518 or fewer, the number of calls in centre B on such days can be as low as 418 whereas the lowest number of calls for centre A is 503. For centre A, on a quarter of the days there are at least 582 calls. For centre B there are at least 601 calls on a quarter of the days. Although the medians are the same for each centre and so are the lower quartiles, the workload is much more variable in centre B than centre A.

<u>Ordinary Certificate, Paper II, 2009. Question 4</u>

The class "30.5 but less than 35.5" has class width 5 (hours).

The area of this bar of the histogram is therefore 5 × 22.

If this is to represent a frequency of 22 then the scale factor is 1/5 = 0.2.

The class "20.5 but less than 30.5" has width 10 (hours).

As the frequency here is 22, we must have Height × 10 × 0.2 = 22, so Height = 11 cm.

Similarly for the "70.5 but less than 90.5" class, the width is 20 (hours) and so we have Height × 20 × 0.2 = 8 so that Height = 2 cm.

We construct a table based on the mid-points of the cells.

Mid-point in hours (x)	Frequency (f)	fx
8.0	6	48
15.5	12	186
25.5	22	561
33.0	22	726
38.0	20	760
45.5	24	1092
60.5	16	968
80.5	8	644
Total	130	4985

$\bar{x}$ = 4985/130 = 38.35 hours, to 2 decimal places.

The table of the upper class boundaries and cumulative frequencies for a cumulative frequency graph is as follows.

Upper class boundary in hours(x)	Cumulative frequency
5.5	0
10.5	6
20.5	18
30.5	40
35.5	62
40.5	82
50.5	106
70.5	122
90.5	130

Ordinary Certificate, Paper II, 2009. Question 5

The table of rankings is as follows.

Country	100m	Hammer	High Jump
GBR	1	7	8
FRA	2	5	4
GER	3	2	1
UKR	4	6	6
POL	5	1	3
GRE	6	3	5
RUS	7	4	2
BEL	8	8	7

The formula for Spearman's rank correlation coefficient is $r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$. Here we have $n = 8$ and therefore $n(n^2 - 1) = 504$.

For 100m and Hammer:

									Total
100m	1	2	3	4	5	6	7	8	
Hammer	7	5	2	6	1	3	4	8	
d	–6	–3	1	–2	4	3	3	0	
d^2	36	9	1	4	16	9	9	0	84

$r_s = 1 - (6 \times 84/504) = 0$

For 100m and High Jump ("Hi Jmp"):

									Total
100m	1	2	3	4	5	6	7	8	
Hi Jmp	8	4	1	6	3	5	2	7	
d	–7	–2	2	–2	2	1	5	1	
d^2	49	4	4	4	4	1	25	1	92

$r_s = 1 - (6 \times 92/504) = 0.0952$

For High Jump ("Hi Jmp") and Hammer:

									Total
Hi Jmp	8	4	1	6	3	5	2	7	
Hammer	7	5	2	6	1	3	4	8	
d	1	–1	–1	0	2	2	–2	–1	
d^2	1	1	1	0	4	4	4	1	16

$r_s = 1 - (6 \times 16/504) = 0.8095$

Solution continued on next page

For 100m and Hammer the value of r_s is 0, indicating that there is no correlation between the rankings in these two events.

For 100m and High Jump the value of r_s is –0.0952, a small negative value. This indicates a slight tendency for a country finishing well in the 100m to do badly in the High Jump and vice versa.

For High Jump and Hammer the value of r_s is 0.8095, a large positive value. This indicates a strong tendency for a country which performs well in the High Jump to also do so in the Hammer and for countries which perform badly to do so in both.

Ordinary Certificate, Paper II, 2009. Question 6

(i) Probability that £250,000 is not in central box in any show = 21/22.

Probability that this happens for 99 shows = $(21/22)^{99}$ = 0.009997 which is slightly less than 0.01.

(ii) (a) If the contestant has a "blue" sum in the central box, there are 10 "blue" and 11 "red" to choose from for the first choice, and so on for the remaining choices.

So the probability of five "blues" is

$$\frac{10}{21}\times\frac{9}{20}\times\frac{8}{19}\times\frac{7}{18}\times\frac{6}{17} = \frac{4}{323} = 0.01238.$$

Similarly, if there is a "red" sum in the central box, there are initially 11 "blue" and 10 "red" to choose from and the probability is

$$\frac{11}{21}\times\frac{10}{20}\times\frac{9}{19}\times\frac{8}{18}\times\frac{7}{17} = \frac{22}{969} = 0.02270.$$

The central box is equally likely to contain a "blue" or a "red", i.e. each has probability ½.

So the overall probability of 5 "blues" is

$$\left(\frac{1}{2}\times\frac{4}{323}\right) + \left(\frac{1}{2}\times\frac{22}{969}\right) = \frac{1}{57} = 0.0175 \text{ to 4 decimal places.}$$

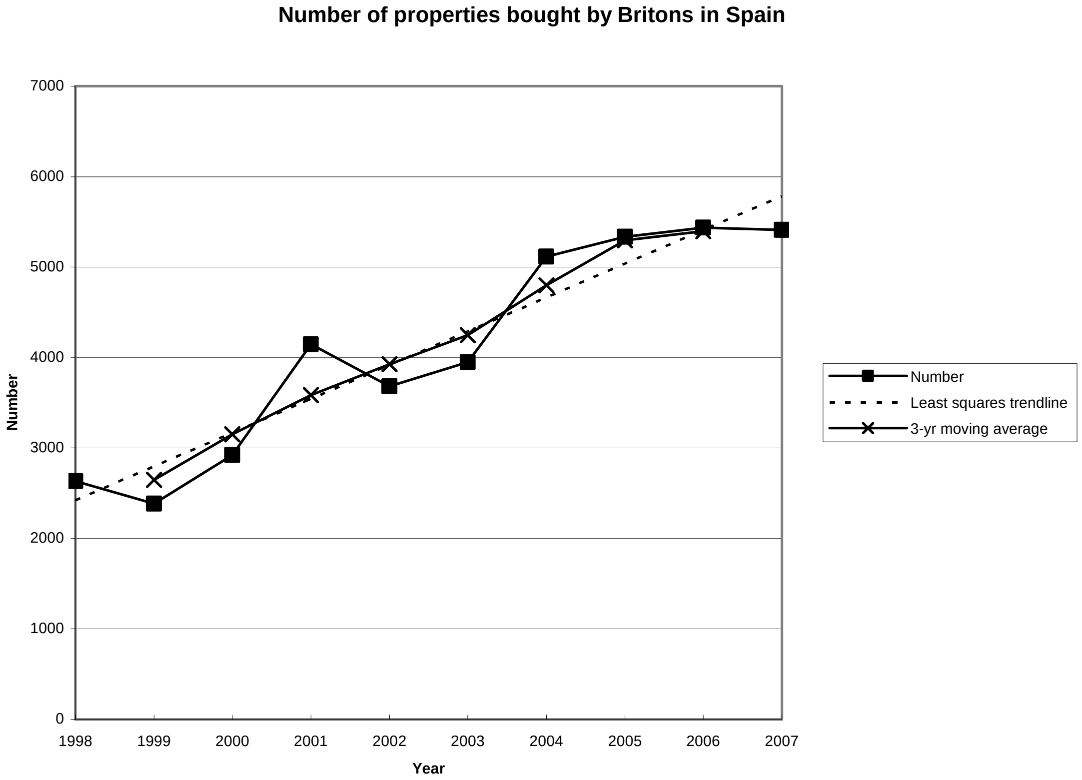
(b) By symmetry, P(5B) = P(5R), P(4B and 1R) = P(1B and 4R), and P(3B and 2R) = P(2B and 3R)

Thus P(5B) + P(4B and 1R) + P(3B and 2R)=0.5

We have that P(5B) = 0.0175 [answer to part (ii)(a)] and P(4B and 1R) = 0.1379 [given in the question].

Thus 0.0175 + 0.1379 + P(3B and 2R) = 0.5, so the probability of 3 "blues" and 2 "reds" is 0.5 – 0.0175 – 0.1379 = 0.3446.

<u>Ordinary Certificate, Paper II, 2009. Question 7</u>



The 3-year moving averages for the number of properties bought by Britons in Spain are as follows. As an example, 2646 is calculated as (2634 + 2385 + 2920)/3 (rounded to the nearest integer for convenience). These have been plotted on the chart above.

Year	Number	3yr moving average
1998	2634	
1999	2385	2646
2000	2920	3151
2001	4148	3584
2002	3683	3926
2003	3947	4248
2004	5114	4799
2005	5336	5296
2006	5438	5395
2007	5412	

The least squares trend line $y = 2421 + 373x$ is plotted on the chart above.

Solution continued on next page

Advantages of moving average:

- Its form is not fixed so its gradient changes as the data change
- On the chart, the gradient is flatter towards the upper end of the series, which seems to match the behaviour of the observed values

Disadvantages of moving average:

- It has no fixed equation; each trend value had to be worked out individually
- Because of the averaging process, there is no value at the start or end of the time period (i.e. here no trend value for 1998 or 2007)

Advantages of least squares trend line:

- It has a fixed equation so its value can be worked out easily at any time point
- To plot the line, we needed to calculate only two values, plus a third one as a check
- It uses values from all the data

Disadvantages of least squares trend line:

- It is always linear regardless of how the data change
- On the chart, the least squares line is still rising at the same rate towards the upper end of the series, although it looks as though the data values are levelling off

We would choose the moving average method, as a linear form for the trend does not appear appropriate.

Ordinary Certificate, Paper II, 2009. Question 8

Company A's share price increased by the greatest amount of money, from 765 to 1566 pence per share, i.e. an increase of 801p.

The price relative for Company A (based on 1 Dec 1998 = 100) is 1566/765 × 100 = 204.7, to 1 decimal place. Similarly for the other companies. Thus we get

Company	Price relative (1 Dec 1998 = 100)
A	205
B	165
C	151
D	456
E	501

Company E's share price has risen by the greatest percentage, 401%.

A weighted index number will allow for the fact that the investor has different values of shareholdings in each company. The greater the value of the shareholding, the greater the investor is affected by changes in the share price.

Value weights are appropriate, i.e. the number of shares multiplied by the price on 1 Dec 2008. The calculation is shown below.

Company	Number of shares held	Price in pence on 1 Dec 1998	Price in pence on 1 Dec 2008	Price relative (1 Dec 1998 = 100)	Weight (p)	Weight × Price relative (p)
A	200	765	1566	205	313200	64206000
B	400	636	1050	165	420000	69300000
C	350	813	1231	151	430850	65058350
D	500	62	283	456	141500	64524000
E	250	153	767	501	191750	96066750
					1497300	359155100

Thus the share price index for 1 Dec 2008 (1 Dec 1998 = 100) is 359155100/1497300 = 239.868 = 240 to the nearest whole number.

The investor's share prices as a whole have risen by 140% from 1 Dec 1998 to 1 Dec 2008.